An Analysis of Topic Modelling for Legislative Texts

James O' Neill*
Insight Centre for Data Analytics
IDA Business Park
Galway, Ireland
james.oneill@insight-centre.org

Cecile Robin†
Insight Centre for Data Analytics
IDA Business Park
Galway, Ireland
cecile.robin@insight-centre.org

Leona O' Brien‡
Governance, Risk and Compliance Technology Centre
University College Cork
Cork, Ireland
leona.obrien@ucc.ie

Paul Buitelaar§
Insight Centre for Data Analytics
IDA Business Park
Galway, Ireland
paul.buitelaar@insight-centre.org

ABSTRACT

The uprise of legislative documents within the past decade has risen dramatically, making it difficult for law practitioners to attend to legislation such as Statutory Instrument orders and Acts. This work focuses on the use of topic models for summarizing and visualizing British legislation, with a view toward easier browsing and identification of salient legal topics and their respective set of topic specific terms. We provide an initial qualitative evaluation from a legal expert on how the models have performed by ranking them for each jurisdiction according to topic coherency and relevance.

CCS CONCEPTS

•**Computing methodologies** → **Non-negative matrix factorization; Latent Dirichlet allocation;** Mixture models;

KEYWORDS

Topic modelling, dimensionality reduction techniques, Bayesian inference, topic coherency

ACM Reference format:
James O' Neill, Cecile Robin, Leona O' Brien, and Paul Buitelaar. 1997. An Analysis of Topic Modelling for Legislative Texts. In *Proceedings of 2nd Work on Automated Semantic Analysis of Information in Legal Texts, London, UK, June 2017 (ASAIL'17),* 8 pages.
DOI: 10.475/123_4

1 INTRODUCTION

The legal domain is experiencing a major shift toward automated tools that can cope with rapid change in the respective legal domains. Regulatory change (RC) is a notable area that has gained much attention in recent years due to the difficulties in compliance. In order to build automated solutions for compliance and verification, automated knowledge acquisition is an imperative for related tasks. An initial step towards such a system requires an overview/summarization of the core topics within the domain, in order to identify salient terms within the topics that are potentially associated with compliance across various documents. Unlike, other approaches that use metadata as a This papers analyzes topic models as well as one domain model to address this. We start by describing the models used in this analysis.

*Carried out the majority of the work
†Supervisory role in the project
‡Provided domain expertise
§Supervisory role in the project

ASAIL'17, London, UK

DOI: 10.475/123_4

2 TOPIC MODELLING

Topic Modelling has many applications in natural language processing [23] and information retrieval, including summarizing the main topics in a domain and ranking documents given a query [29], the former in which we are interested in.

2.1 Dimensionality Reduction Approaches

A basic approach to modeling topics is to view a corpus as a set of term frequencies (tf) where the weight for each term is also dependent on the inverse document frequency (idf), as shown in equation 1 (e.g "and" occurs many times in a document, therefore its weight is low). Here N represents the number of documents and n_t is the number of documents term t appears in.

$$f_{t,d} * \log \frac{N}{n_t} \tag{1}$$

From a term-document matrix M, dimensionality reduction techniques are often used to reduce all terms to a set of concepts, which can be interpreted as approximations of "topics" in a given corpus. The matrix factorization techniques we discuss include Singular Value Decomposition (SVD) and Non-Negative Matrix Factorization (NMF).

2.1.1 Non-Negative Matrix Factorization. NMF is specifically for factorizing matrices with non-negative values, hence why it is particularly suitable for term-document matrices. Since M is represented as non-negative values, features are composed of additive computations resulting in a part based representation (as opposed to subtracting values which would not lead to parts-based factored representation) [7].

The objective of NMF is to find an approximation of matrix M by factorizing it into $W(r\ x\ k)$ and $H(k\ x\ c)$ such that $M \approx WH$ and k have lower rank[1] than M. The reconstruction error is minimized according to that shown in equation 2 [14, 15].

$$\frac{1}{2}|M - WH|_F^2 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}(M_{ij} - WH_{ij})^2 \tag{2}$$

Also described by Lee and Seung [14], the multiplicative update algorithm is used for updating both W and H. Both update rules are outlined in equation 3. The equations ensure the minimization is constrained to W and H being positive and also that the distance D between both is positive.

$$\begin{aligned} H_{\alpha,\mu} &:= H_{\alpha,\mu}\frac{(W^T M)_{\alpha,\mu}}{(W^T WH)_{\alpha,\mu}}, \\ W_{i,\alpha} &:= W_{i,\alpha}\frac{(MH^T)_{\alpha,i}}{(WHH^T)_{\alpha,i}} \end{aligned} \tag{3}$$

In this work, instead of using gradient descent to minimize the sum of squared (euclidean) distance (SSD) between M and WH, we use the Coordinate Descent solver, provided in *scikit-learn* [20]. Lin et al. [16] describe the process that builds upon the multiplicative update algorithm by applying Alternating Non-negative Least Squares[2] (ANLS) using projected gradient descent which is a parameter estimator with lower-bounded constraints. Although, NMF is widely used for topic modeling [31], it is sometimes known to produce non-meaningful topics, particularly if a term-document matrix is relatively sparse. Therefore, the identification of both rare and non-distinct terms is an important step to consider for removal before factorization. Furthermore, NMF can be prone to local minima[3].

2.1.2 Singular Value Decomposition. SVD decomposes a matrix into three parts as shown in equation (4) in order to find a lower rank[4] approximation of the term-document matrix. Consider M to be a tf-idf matrix representation of the corpus, where U diagonalizes MM^T and u_i represents the corresponding eigenvector. Similarly V^*[5] diagonalizes $M^T M$ and v_i represents $M^T M$ eigenvectors. The diagonal values of Σ are ordered singular values[6].

$$M = U\Sigma V^* \tag{4}$$

SVD on a term-document matrix is also referred to as Latent Semantic Analysis (LSA), as the lower ranked matrix M is said to represent a latent semantic space. In information retrieval, it is referred to as Latent Semantic Indexing (LSI), where SVD is used to index documents by representing documents (document-document) and terms (document-term where terms are query terms) in vector space where the elements in the vector correspond to the degree that a term or document has to a given topic. The similarity between a query and a given set of documents can then be determined using a term-topic-matrix [24]. This is particularly helpful for distinguishing polysemous and synonymous terms.

2.2 Probabilistic Topic Modelling

Although, both NMF and SVD are well-founded dimensionality reduction techniques they do not provide the flexibility of a probabilistic framework that can introduce priors. Hence we introduce a set of probabilistic based topic models.

2.2.1 Probabilistic Latent Semantic Analysis. In contrast with LSA, probabilistic Latent Semantic Indexing (pLSA) is an aspect model (where aspects are unobserved latent variables) instead of a dimensionality reduction technique such as SVD in LSA. In this sense, a term t with a document d in a term-document matrix can be described as being generated by a joint distribution $P(t, d)$. PLSA proposes to use maximum likelihood estimation (MLE) to estimate the probability of terms in a collection D given a set of parameters θ and topics z as shown in equation 5.

$$P(D|\theta) = \sum_{k=1}^{K} P(D|\theta, z_k)P(z_k|\theta) \tag{5}$$

The parameters for the equation 6 are $P(t|z_k)$ and $P(z_k|d)$ for computing conditional probability $p(t|d)$ and requires a product of sums $P(d)$.

$$\begin{aligned} P(t|d) &= \sum_{k=1}^{K} P(t|z_k)P(z_k|d), \\ P(d) &= \prod_{t=1}^{T}\sum_{k=1}^{K} P(t|z_k)P(z_k|d)^{X(t,d)} \end{aligned} \tag{6}$$

Obtaining the optimal parameters for both requires that the likelihood is not directly computed but rather the log likelihood, with the constraint that the probability $P(t|k)$ for all t must sum to 1 and $P(z_k|d)$ for all topics z_k must sum to 1 as shown in 7[7].

$$\log P(d) = \sum_{d=1}^{N}\sum_{t=1}^{T} X(t,d)\log\sum_{k=1}^{K} P(t|z_k)P(z_k|d) \tag{7}$$

The KL divergence[8] is then used in equation 8 to minimize the distance between the estimated topic distribution $q(z)$ and $P(z|X, \theta)$. Since obtaining MLE directly is difficult, a lower bound is often placed on the likelihood estimate in practice. Expectation Maximization (EM) is often used to optimize equation 8, a common algorithm for maximizing the likelihood for models with latent variables such as topics Z.

$$\log P(X|\theta) - KL(q(z)||P(z|X,\theta)) \tag{8}$$

In summary, pLSA has advantages over previously mentioned NMF and LSA topic models by modeling topic distributions for specific documents in a generative modeling procedure. Although,

[1] The rank of a matrix is the amount of linearly independent column or row vectors
[2] see [16] for more technical detail
[3] A state where the model converges to a non-global solution, that can lead to poor parameter estimates.
[4] The rank of a matrix is the number of linearly independent column vectors in a matrix (e.g document-term matrix), which can be used to reconstruct all column vectors.
[5] The * denotes the transpose of a conjugate matrix in the case that the coefficients are complex. If real valued, it simply represents a matrix transpose.
[6] singular values are the square root of the eigenvalue
[7] There is a number of algebraic steps in solving for this equation, for the interested reader please refer to the following tutorial [13]
[8] KL divergence measures the distance between two distributions

all three models neglect word and document ordering, following the "exchangeability" assumption [2], hence there is no distribution over the mixture of topics, which commonly results in overfitting. One model that does allow for this Bayesian approach to estimation instead of maximum likelihood is Latent Dirichlet Allocation (LDA).

2.2.2 Latent Dirichlet Allocation. Latent Dirichlet Allocation was first introduced by Blei et al. [3] and has since been a state of the art (SoTA) topic model, showing to have more expressiveness over probabilistic LSA (pLSA) [4, 9]. LDA builds a Bayesian generative model using Dirichlet priors for topic mixtures (an assumed prior probability for each topic distribution, Dirichlet is a set of categorical distributions in this sense), in contrast to pLSA that can be considered to use uniform prior distribution for the topic mixtures. Further extensions since then have been made to improve and adapt this model in a continuous space setting. In this sense, continuous word embeddings are used. Categorical distributions are replaced with multivariate Gaussian distributions, meaning that Gaussian LDA has the capability of handling out of vocabulary words on unseen text [10]. Figure 1 illustrates LDA in a block diagram. Here the probability of word w is dependent on a topic k in z which is dependent on probability of a document θ_d that is drawn from a Dirichlet prior α. Likewise the a word w is also dependent on the probability ϕ that a word w is in topic k

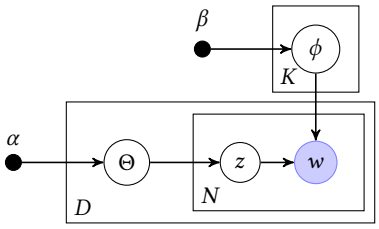

Figure 1: Latent Dirichlet Allocation

The LDA generative process is described in algorithm 2.2.2 as described by Blei [4]. . Here for each document a parameter θ_d is chosen from a Dirichlet prior distribution, then for each word in d a topic category is chosen according to the Dirichlet. A word w is then generated given the topic z_w and β.

for document d in corpus D **do**
  Choose $\theta_d \sim$ Dirichlet(α)
  for position w in d **do**
    Choose a topic $z_w \sim Categorical(\theta_d)$
    Choose a word w_w from $p(w_w|z_w, \beta)$, a categorical distribution over words conditioned on the topic and the prior β.
  end for
end for

The aforementioned Gaussian LDA represents these words as continuous embedded vectors instead of discrete co-occurrence counts, replacing the categorical distributions for z_n and w_n with Gaussians.

The saliency of terms within a topic is considered by [8] and formulated in equation 9. A distinctive word w is a word that has a higher log-likelihood of being in a topic K compared to a random word. Hence, if a word w occurs many topics it is non-informative, resulting in lower saliency. The more informative topic specific terms and less general terms are desired by the legal practitioners, hence we use this saliency measure in our analysis.

$$S(w) = P(w) \sum_{T} P(K|w) \log \frac{P(K|w)}{P(K)} \tag{9}$$

Sievert and Shirley [25] describe the relevance measure, also shown in 10, where $\phi_{k(w)}$ is the probability of w for topic k and $p(w)$ is the probability of observing w in corpus D. In this work, λ is set to multiple settings, in which λ will be set according to relevance of terms a legal practitioner has informed us of, prior to the final analysis of each topic model. As mentioned by [25], ϕ is commonly estimated using Variational Bayes (VB), similarly we employ both VB for posterior estimation.

$$r(w, k|\lambda) = \lambda \log(\phi_{k(w)}) + (1 - \lambda) \log(\frac{\phi_{k(w)}}{p(w)}) \tag{10}$$

2.2.3 Hierarchical Dirichlet Process. The hierarchical Dirichlet process (HDP) is a parameter free model, in contrast to the standard LDA algorithm. The number of topics are inferred from the corpus and not chosen beforehand, as is the case with LDA. HDP requires a two-stage process where the common core distribution represents a Dirichlet process (DP) where DPs all share this common distribution. The common corpus distribution is sampled from for each cohort of data instances as described by Wang et al. [27], shown in equation 11. C_0 is the core Dirichlet distribution, H is a Dirichlet distribution, α is the inverse variance of the H DP ($\frac{H(A)(1-H(A))}{\alpha+1}$, A being a subset of the corpus) and index j is an iteration over each DP subset.

$$\begin{aligned} C_0 \; & DP(\gamma H) \\ C_j \; & DP(\alpha, C_0) \end{aligned} \tag{11}$$

Wang et al. [27] propose the use of a computationally feasible online variational Bayesian inference for this task, similarly we follow the same approach using the *gensim*[9] library. Variational Bayes is a variational method that tries to find an approximate joint distribution with latent variables $Q(x; \theta)$ such that this approximate distribution is close to the true distribution $P(x)$. To minimize the asymmetric distance between both distributions KL-divergence[10] is used $KL(Q(x; \theta)||P(x))$.

HDP has the advantage over previously mentioned topic models in that the number of topics do not have to be chosen apriori to modeling. This practical advantage is useful in the context of this work as the number of topics emerging can vary (depending on the jurisdiction, publication year etc.) and are not explicitly known. The main disadvantage over LDA from an implementation perspective is that it is more complicated to implement and requires extra computation to use variational Bayes.

[9] see here-https://radimrehurek.com/gensim/models/hdpmodel.html
[10] Measures the distance between two distributions.

2.3 Saffron

Saffron is a software tool[11] that can construct a model-free topic hierarchy. It extracts topics related to the domain of expertise, along with the semantic relatedness of the terms between them, and constructs a taxonomy out of it. *Saffron* can deal with multiword expressions (as opposed to single terms of that proposed in the previously mentioned topic models) with the goal of improving the topic coherency within the domain description.

2.3.1 Saffron Domain Modelling. Saffron builds this topic hierarchy from a corpus using feature selection during a term and linguistic pattern extraction phase. Mimno [18] discuss the importance of topic coherency, which is important issue for statistically driven models in order for SMEs to rely upon them. This issue is addressed with *Saffron* by using the following constraints: only extracting contentful parts-of-speech only (that is noun, verbs and adjectives), limited multiword expression length (very long expressions are too specific while shorter ones are more generic) and the distribution of a term across the corpus (generic words of an area of expertise are meant to appear in at least 1/4 of the documents of the corpus).

2.3.2 Saffron Topic Extraction. The domain hierarchy is then used as a base for topic extraction, in order to measure a terms coherence within the domain. It features techniques using context words: lexical coherence of candidates with general terms, domain coherence calculated using Pointwise Mutual Information (PMI), use of top level terms to extract intermediate terms, etc.

2.3.3 Saffron Taxonomy Construction. To create the pruned graph which represents the taxonomy, the measure of the strength of the relationships between two research terms is calculated, defined as follow[12]:

$$I_{ij} = \frac{D_{ij}}{D_i * D_j} \tag{12}$$

Edges are added in the graph for all the pairs that appear together in at least three documents, threshold fixed based on the results of previous studies and tests (see [5] for more details). Saffron also uses a generality measure to direct edges from generic concepts to more specific ones. This results in a dense, noisy directed graph that is further trimmed using a specific branching algorithm which was successfully applied for the construction of domain taxonomies in [19]. This yields a tree structure where the root is the most generic term and the leaves are the most specific terms.

Saffron Graph Visualization. To visualize the obtained topical hierarchy, *Saffron* incorporates Cytoscapes graph visualization tool (see http://www.cytoscape.org/) to perform a network analysis and a customization of the output layout. In our layout, the size and the colour of the nodes are proportional to the number of neighbours each topic connects to and different colours represent the level in the hierarchy, red being the root, orange a child node of the root and turquoise nodes are parents of the leaf nodes that are blue.

[11] see here - http://saffron.insight-centre.org/

[12] where D_i is number of articles that mention the term Ti in our corpus, D_j is number of articles that mention the term T_j, and $D_i j$ is the number of documents in which both terms appear

3 RELATED WORK

Wiltshire et al. [30] introduced a large scale machine learning systems that incorporates the use of hierarchical topic construction after the extraction of terms, legal phrases and case cites. Their system allows for a ranking and classification of topics given a legal concept as input according to a scoring criterion. George et al. [12] provide a legal system for ranking documents according to their similarity to legal cases by finding similarity between documents in the latent topic space and query terms. They then use human assistance to provide annotate documents that are relevant to the query in a semi-supervised fashion. In contrast, our work is fully unsupervised with no human assistance during the topic modelling process. LDA has been used extensively on natural language texts such as social media texts [21], publication texts, newspapers etc. and typically not in formal settings such as their use on legal texts.

Raghuveer and Kumar [22] use LDA to cluster Indian legal judgements and use cosine similarity as the distance measure between documents for clustering. However, their evaluation does not present the prior knowledge of a legal expert to determine if the clusters coincide with legal knowledge within the domain.

Wang [28] apply the nonparametric Bayesian HDP to tractography segmentation (segments images of neural tracts that use techniques for Magnetic Resonance Imaging (MRI)) to identify clusters of fibres in the brain. Teh et al. [26] use HDP on biology abstracts, NIPs (Neural Information Processing Systems) sections and the book "Alice in Wonderland", three quite distinct text registers.

In all texts, HDP automatically inferred approximately the same number of topics as LDA after LDA underwent model selection according to the log-likelihood of the topic distributions.

Ahmed and Xing [1] use dynamic HDP to track topic over time, documents can be exchanged however the ordering is intact. They also use longitudinal NIPS papers to track emerging topics and decaying topics (this is worth noting, particularly for tracking changing topics around compliance issues).

The use of the aforementioned *Saffron* has been previously demonstrated through a wide range of projects from several domains and for different tasks. In [6], Bordea used Saffron's topic extractor to analyze legal documents arising around the financial crisis in 2008. She mapped the problem as an expert finding task, which aims at ranking people that have knowledge about a given topic. In that particular context, the task allowed the identification of individuals involved in defining the response of the U.S. government to the financial crisis by searching for a topic of interest. In [5], Saffron was used as a tool to detect the presence of different disciplines within the field of Web Science. By running it on over 10 years of Web Science conference series documents, it resulted on a discovery of four communities (Communication, Computer Science, Psychology, and Sociology), and trends over time and types of paper. Saffron was also used in a demo for an Irish bookshop website[13] to extract topics from book descriptions/reviews and then classify them accordingly. It was also used to link the books for the creation of a multi-level browsing application for book navigation.

Lu et al. [17] cluster documents with topic segmentation on a large collection of legal documents including " judicial opinions,

[13] see http://kennys.insight-centre.org/

statutes, regulations, administrative materials and analytical documents". Legal experts first predefine the topics according to their expertise which is used as a baseline to compare the quality of the topics extracted from the soft clustering algorithms. The pipeline involves the topic clustering of segmented documents from metadata. A large set of topics are then merged using an SVM ranker. These merged clusters are then associated with documents to produce a document-topic segmentation. To our knowledge, this work is the most relevant application of topic clustering for legal documents. Although, their pipeline uses metadata during the topic clustering process which is not always available and as to this date there is not an agreement within the legal domain upon a standard schema for legal documents. Although, *LegalDocML* (previously known as AkomoNtoso XML) has made a significant effort to overcome such issues, it is still in its infancy and thus far, it is not guaranteed to adapt to changes over time. Therefore, our focus is on raw text and assumes no use of a markup language. Hence, we introduce the methodology from dataset description and text processing to the modelling procedure and legal experts evaluation.

4 METHODOLOGY

This section outlines the steps towards creating each topic model and their configurations used for analysis. We start with a brief introduction to the corpora used and preprocessing steps common to all topic models.

4.1 Dataset Description

United Kingdom legislative texts were used for topic modeling[14]. The corpus contains 41,518 documents between 2000 - 2016. However, for practical purposes the analysis is carried out on the year 2016 only to lessen the reading burden on the legal practitioner. The legislative types consist of the following: 304 *Northern Ireland Statutory Rules*, 838 *UK Statutory Instruments*, 132 *Welsh Statutory Instruments* and 317 *Scottish Statutory Instruments*.

4.2 Text Preprocessing

Corpus specific regular expressions (RE) are used to clean legal domain syntax (e.g bracketed alphanumerics), followed by tokenization and lemmatization using the WordNet lemmatizer [11]. The structure usually contains nested expressions e.g (ii) followed by (a) and (b) subsections. This syntax is removed using the regular expressions along with other standard RE for identifying references and alphanumeric expressions e.g "*Regulation EC No. 1370/2007* means Regulation 1370/2007 ..." Stopwords are used to remove redundant and non-consequential terms in the corpora, by removing words that have a frequency $f < 2$. This is carried out under the superivsion of a subject expert by analysing a subsample of terms which are considered for removal. We assume that terms with high frequency are not specific to a particular topic e.g 'the','of' etc. Also, rare terms that occur infrequently are not representative of a single topic since they do not appear enough to infer that it is salient for a topic. Each corpus (corpus per jurisdiction) is then converted to a term-document matrix where weights are placed on each word using

[14]Retrieved from: http://www.legislation.gov.uk/

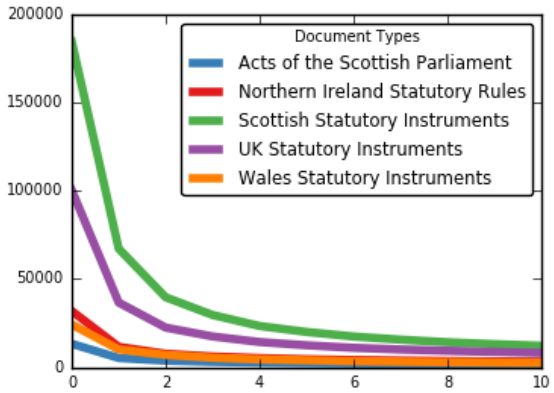

Figure 2: Rare-word Removal For Each Corpus

the aforementioned tf-idf weighting scheme. Furthermore, 30 terms for all models except Saffron are listed for an SME for ranking. For *Saffron* we rely on a visualization of the term hierarchy for a domain expert to judge.

4.3 Model Configurations

This section details the topic model parameters. Since the term-document matrix is quite sparse (evident from 2), NMF is initialized using Non-Negative Singular Value Decomposition (NNSVD). The Coordinate Descent solver is used for minimizing the reconstruction error as mentioned in section 2.1.1. The number of components is set to $n_k = 10$. LSI uses standard SVD which does not require much tuning only to choose the number of singular values, also set $n_k = 10$. For LDA we choose low relevance $\lambda = 0.25$ to highlight topic specific terms. As mentioned, HDP infers the number of topics therefore we

4.4 Ranking Criterion

In order for a legal practitioner to assess the models in a fair manner, a set of guidelines are presented for the ranking of the models. An important aspect to ranking is the pretuning of the term relevance parameter λ, which chooses the top 30 terms that are presented for each topic within the jurisdiction accordingly. We also assess a number of parameter setting for NMF, LSA, LDA and HDP before finally choosing the final 10 set of topics which the legal expert makes their final judgment.

5 RESULTS

In this section we analyse and the topics retrieved for each approach and SME evaluated topics for the regulations. Figure 2 simply compares the effects of dictionary size once infrequent terms are increasingly removed. It is evident that after removing terms that occur less than twice the corpus size dramatically decreases, meaning that a significant number of terms are too specific too a particular document. We remove these terms for subsequent analysis.

Latent Dirichlet Allocation Visualization. For visualization of LDA topics we use the *pyLDAvis* [25] visualization tool. Multidimensional scaling is used to project the t dimensional space to a 2 dimensions as shown in figure 3. Ten topics for Northern Ireland Statutory Rules (NISR) are presented with the relevance metric set $\lambda = 0.25$ (decides the term-topic specificity). This is done under the supervision of a legal practitioner to ensure λ is tuned to correct specificity and topics that are also coherent before a final evaluation.

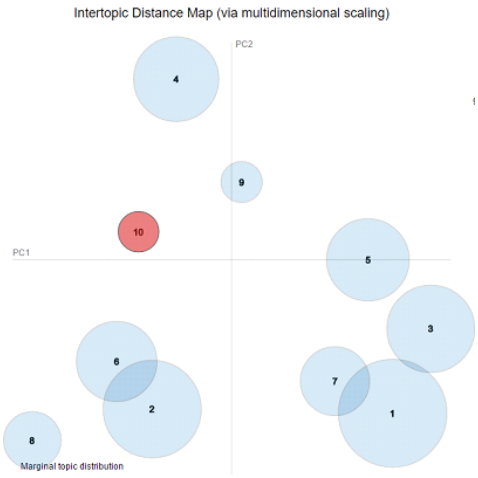

Figure 3: Latent Dirichlet Allocation topics for Northern Ireland Statutory Rules projected to 2 principal components using multi-dimensional scaling (MDS)

Some terms such as *biomass*, *biomaterial*, *bioliquid*, *fossil* and *fuel* show a clear and distinct topic and are quite topic specific given $\lambda = 0.25$, shown by red bars which indicate the term frequency with the given topic as opposed to the blue bar that indicate the term frequency among the whole corpus.

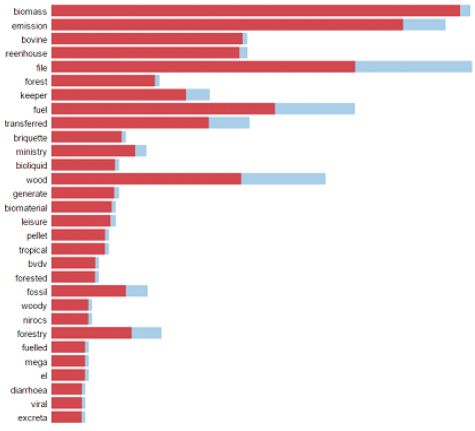

Figure 4: Latent Dirichlet Allocation terms for topic 10 of Northern Ireland Statutory Rules

Saffron. Figure 5 illustrates *Saffron* for Northern Ireland Statutory Rules. As expected, the most prominent nodes include high-level entities such as *european union*,*statutory rules*, *northern ireland*, *united kingdom*, *floating structure*, *support allowance*,*food safety* and *police service* all emerge as topical noun phrases. The root of the graph is without surprise the term *northern ireland*, which also has the highest degree (number of incoming and outgoing links). Note that *northern ireland* can embody different roles/facets: it can refer to the political entity, the administrative state, the province, the legal jurisdiction, etc., which can result in many different possible connections and clusters of a different sense. We notice here a few different topics which Northern Ireland is connected to. Apart from isolated nodes already referring to different facets of this entity (like public service ombudsman, regional development, civil legal service, etc.), two main clusters can be identified deriving from the root.

We locate a cluster around the extracted topic of the *department of justice*, and one of the *support allowance*, the latter from which derives the whole graph of topics for the Northern Ireland Statutory Rules. This topic thus seems to be the main focus of the content of this corpus for 2016. In figure 6 we analyze a subset of this graph (and thus sub-domains) that includes *housing benefit*, *income support*, *social security*, *personal˙independence payment*. They are all semantically related to the mother node *support allowance*, but tackling different aspects of it. We can see the advantage of the hierarchical structure of the graph, with semantically related topics that go from the more generic to the more specialized one. This way, we identify a waterfall structure from the *housing benefit* branch, which is logically followed by the more specific *local housing allowance*, and then *local housing allowance determination*. Another quite clear example can be observed from the *child support* branch, related to the *personal independence˙payment* node. From *child support*, the directed edge goes to *child support maintenance*, then *maintenance calculation*, and finally the three topics *child˙support˙maintenance˙calculation˙regulation*, *welfare service* and *maintenance assessment*. The *police service* node is the root of a taxonomy that includes children nodes *northern˙ireland˙reserve* $\Rightarrow$ *notice˙of˙appeal* $\Rightarrow$ *written˙representation,avoiding service* $\Rightarrow$ *reasonable˙amount˙of˙duty˙time*. This example summary allows a legal practitioner to identify topics surrounding certain legal issues or for simply summarizing a complete jurisdiction. Zooming in on a subset of the hierarchical tree we highlight a topic and coherent multi-word expressions summarizing an area within Northern Ireland Statutory Rules in figure 7.

Ranking. Table 1 shows the results of SME ranking after assessing each topic model for each jurisdiction. Saffron overall is favored for all jurisdictions, considering it is the only model that performs multi-word expression topic extraction and weighting of descriptive noun terms/phrases. We conjecture that the appeal of a hierarchical structure and multi-word noun expressions has influenced the interpretation of the salient terms in the domain, making it easier for legal practitioners to identify important and coherent legal topics.

We emphasize at this point that single word topic models and multi-word hierarchical models are not directly comparable for this reasons outlined however, they are included in table 1 to highlight the importance of longer expressions that are linked in a taxonomy,

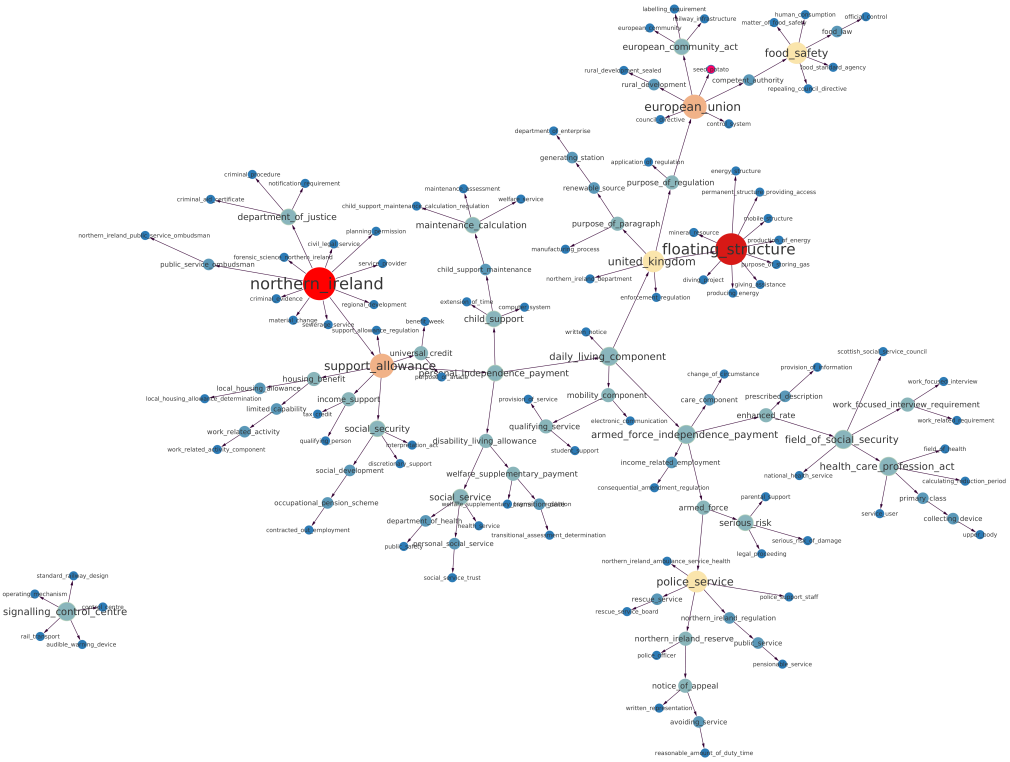

Figure 5: Saffron Hierarchical Topic Extraction for Northern Ireland Statutory Rules

Table 1: Subject Matter Expert Ranking of Topic Models

Rank	Northern Ireland Statutory Rules	Scottish Statutory Instruments	UK Statutory Instruments	Welsh Statutory Instruments
1	Saffron	Saffron	Saffron	Saffron
2	LDA	LDA	LDA	LDA
3	HDP	NMF	HLDP/LSI	HLDP/LSI
4	LSA	LSI	HLDP/LSI	HLDP/LSI
5	NMF	HLDP	NMF	NMF

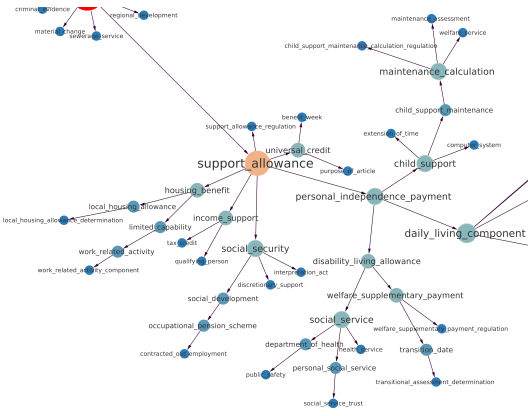

Figure 6: Support Allowance topic within Northern Ireland Statutory Rules

providing more clarity on what the emerging topics are in legal corpus.

6 CONCLUSION

This work has presented a fully automated approach for identifying topics in regulations that assist in easier tracking of important domain terms that correspond to compliance related issues. After evaluation *Saffron* has been consistently ranked as the most favourable of all models, as the aforementioned vocabulary pruning and usage of multi-word expressions has played a fundamental role in topic coherency. Standard LDA has performed the best of all single term models, particularly when top terms are chosen according to their topic specificity. HDP has inferred a similar number of topics as that of LDA according to an analysis of the log-likelihood curve and the legal practitioners judgment. This work is an early indication as to how legal practitioners can identify salient and coherent topics using automatic topic modelling tools.

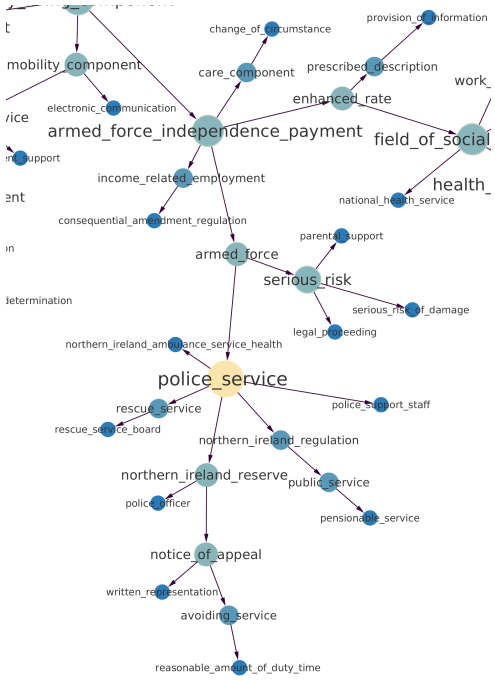

Figure 7: Police Service topic within Northern Ireland Statutory Rules

7 FUTURE WORK

In the future we plan on collaborating with a number of legal practitioners that are experts in specific jurisdictions. This will allow a more generalized overview of the field and subsequently a larger sample of practitioners for the evaluation of these models presented.

REFERENCES

[1] Amr Ahmed and Eric P. Xing. 2012. Timeline: A Dynamic Hierarchical Dirichlet Process Model for Recovering Birth/Death and Evolution of Topics in Text Stream. *CoRR* abs/1203.3463 (2012). http://arxiv.org/abs/1203.3463

[2] David J Aldous. 1985. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII1983*. Springer, 1–198.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3 (March 2003), 993–1022. http://dl.acm.org/citation.cfm?id=944919.944937

[5] Georgeta Bordea. 2013. *Domain adaptive extraction of topical hierarchies for Expertise Mining*. Ph.D. Dissertation.

[6] Georgeta Bordea, Kartik Asooja, Paul Buitelaar, and Leona OBrien. 2014. Gaining insights into the Global Financial Crisis using Saffron. *NLP Unshared Task in PoliInformatics* (2014).

[7] Deng Cai, Xiaofei He, Xiaoyun Wu, and Jiawei Han. 2008. Non-negative matrix factorization on manifold. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE, 63–72.

[8] Jason Chuang, Christopher D Manning, and Jeffrey Heer. 2012. Termite: Visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*. ACM, 74–77.

[9] Patricia J Crossno, Andrew T Wilson, Timothy M Shead, and Daniel M Dunlavy. 2011. Topicview: Visually comparing topic models of text collections. In *Tools with Artificial Intelligence (ICTAI), 2011 23rd IEEE International Conference on*. IEEE, 936–943.

[10] Rajarshi Das, Manzil Zaheer, and Chris Dyer. 2015. Gaussian LDA for Topic Models with Word Embeddings.. In *ACL (1)*. 795–804.

[11] Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

[12] Clint Pazhayidam George, Sahil Puri, Daisy Zhe Wang, Joseph N Wilson, and William F Hamilton. 2014. SMART Electronic Legal Discovery Via Topic Modeling.. In *FLAIRS Conference*.

[13] Liangjie Hong. 2012. A Tutorial on Probabilistic Latent Semantic Analysis. Cornell University Library arXiv, 1–7.

[14] Daniel D Lee and H Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 6755 (1999), 788–791.

[15] Daniel D Lee and H Sebastian Seung. 2001. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*. 556–562.

[16] Chih-Jen Lin. 2007. Projected gradient methods for nonnegative matrix factorization. *Neural computation* 19, 10 (2007), 2756–2779.

[17] Qiang Lu, Jack G. Conrad, Khalid Al-Kofahi, and William Keenan. 2011. Legal Document Clustering with Built-in Topic Segmentation. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM '11)*. ACM, New York, NY, USA, 383–392. DOI:http://dx.doi.org/10.1145/2063576.2063636

[18] David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 262–272.

[19] Roberto Navigli, Paola Velardi, and Stefano Faralli. 2011. A Graph-based Algorithm for Inducing Lexical Taxonomies from Scratch. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three (IJCAI'11)*. AAAI Press, 1872–1877. DOI:http://dx.doi.org/10.5591/978-1-57735-516-8/IJCAI11-313

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[21] Marco Pennacchiotti and Siva Gurumurthy. 2011. Investigating topic models for social media user recommendation. In *Proceedings of the 20th international conference companion on World wide web*. ACM, 101–102.

[22] K Raghuveer. 2012. Legal documents clustering using latent dirichlet allocation. *IAES Int. J. Artif. Intell.* 2, 1 (2012), 34–37.

[23] Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer.

[24] Barbara Rosario. 2000. Latent semantic indexing: An overview. *Techn. rep. INFOSYS* 240 (2000).

[25] Carson Sievert and Kenneth E Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*. 63–70.

[26] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2004. Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes.. In *NIPS*. 1385–1392.

[27] Chong Wang, John Paisley, and David M Blei. Online Variational Inference for the Hierarchical Dirichlet Process. (????).

[28] Xiaogang Wang, W Eric L Grimson, and Carl-Fredrik Westin. 2011. Tractography segmentation using a hierarchical Dirichlet processes mixture model. *NeuroImage* 54, 1 (2011), 290–302.

[29] Xing Wei and W Bruce Croft. 2006. LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 178–185.

[30] James S Wiltshire Jr, John T Morelock, Timothy L Humphrey, X Allan Lu, James M Peck, and Salahuddin Ahmed. 2002. System and method for classifying legal concepts using legal topic scheme. (Dec. 31 2002). US Patent 6,502,081.

[31] Xiaohui Yan, Jiafeng Guo, Shenghua Liu, Xueqi Cheng, and Yanfeng Wang. 2013. Learning topics in short texts by non-negative matrix factorization on term correlation matrix. In *Proceedings of the 2013 SIAM International Conference on Data Mining*. SIAM, 749–757.